Learning Quality-Aware Representation for Multi-Person Pose Regression

Yabo Xiao,[1,*] **Dongdong Yu,** [2,*] **Xiao Juan Wang,** [1,†] **Lei Jin,** [1] **Guoli Wang,** [3] **Qian Zhang** [4]

[1] Beijing University of Posts and Telecommunications
[2] ByteDance Inc.
[3] Tsinghua University
[4] Horizon Robotics

{xiaoyabo, wj2718, jinlei}@bupt.edu.cn, yudongdong@bytedance.com, wangguoli1990@mail.tsinghua.edu.cn,
qian01.zhang@horizon.ai

Abstract

Off-the-shelf single-stage multi-person pose regression methods generally leverage the instance score (i.e., confidence of the instance localization) to indicate the pose quality for selecting the pose candidates. We consider that there are two gaps involved in existing paradigm: 1) The instance score is not well interrelated with the pose regression quality. 2) The instance feature representation, which is used for predicting the instance score, does not explicitly encode the structural pose information to predict the reasonable score that represents pose regression quality. To address the aforementioned issues, we propose to learn the pose regression quality-aware representation. Concretely, for the first gap, instead of using the previous instance confidence label (e.g., discrete {1,0} or Gaussian representation) to denote the position and confidence for person instance, we firstly introduce the Consistent Instance Representation (CIR) that unifies the pose regression quality score of instance and the confidence of background into a pixel-wise score map to calibrates the inconsistency between instance score and pose regression quality. To fill the second gap, we further present the Query Encoding Module (QEM) including the Keypoint Query Encoding (KQE) to encode the positional and semantic information for each keypoint and the Pose Query Encoding (PQE) which explicitly encodes the predicted structural pose information to better fit the Consistent Instance Representation (CIR). By using the proposed components, we significantly alleviate the above gaps. Our method outperforms previous single-stage regression-based even bottom-up methods and achieves the state-of-the-art result of 71.7 AP on MS COCO test-dev set.

Introduction

Given an input RGB image, multi-person pose estimation aims to detect the keypoint positions for all persons. With the prevalence of deep learning (Deng et al. 2009; Newell, Yang, and Deng 2016; Ren et al. 2015), it has attracted much attention since it plays as an important role in many computer vision tasks such as pose tracking (Xiao, Wu, and Wei 2018; Yu et al. 2018a), activity recognition (Li et al. 2019; Shi et al. 2019), human re-identification and so on.

Most existing multi-person pose estimation methods follow top-down pipeline (Chen et al. 2018; Sun et al. 2018;
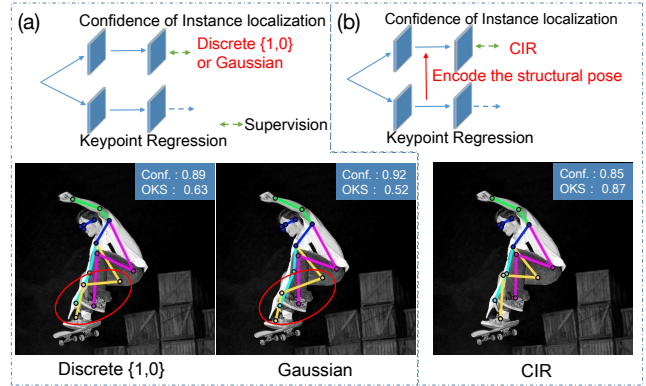

Figure 1: (a) Off-the-shelf single-stage pose regression pipeline and the predicted human pose with high instance score (Conf.) and inferior pose quality (wrong localization for leg area) when using discrete {1,0} or Gaussian kernel to supervise the instance score. (b) Our proposed pipeline predicts the pose with high instance score (Conf.) and superior pose regression quality. The pose regression quality is quantified as OKS.

Yu et al. 2018b; Xiao, Wu, and Wei 2018; Su et al. 2019) and bottom-up pipeline (Cao et al. 2017; Cheng et al. 2020; Kreiss, Bertoni, and Alahi 2019; Papandreou et al. 2018; Newell, Huang, and Deng 2017). The top-down methods firstly detect the region of person instance via object detector (Cai and Vasconcelos 2018; Law and Deng 2018; Tian et al. 2019), then perform single person pose estimation on the cropped human body regions. Generally, the top-down pipeline is limited by the detection-first paradigm which leads to high computation and memory cost. The bottom-up methods firstly locate the keypoints of all persons in an image simultaneously and then assign the keypoints to individuals via a grouping process. However, the additional grouping process is computationally complex.

By contrast, recent practices (Nie et al. 2019; Tian, Chen, and Shen 2019) deliver a single-stage solution via pixel-wise keypoint regression, which is more straightforward and bypasses the above drawbacks of previous methods. In particular, it decomposes the pose estimation task into person instance localization and corresponding keypoint regression.

For example, SPM (Nie et al. 2019) leverages the Gaussian kernel to indicate the confidence of instance localization and proposes a hierarchical pose representation to regress joints. DirectPose (Tian, Chen, and Shen 2019) employs the discrete instance confidence label {1,0} to denote the position and confidence of the person instance and presents a KPAlign scheme to locate the keypoints. However, we consider that there are two gaps in the above practices: 1) As shown in Figure 1(a), the instance score supervised by prior instance confidence label (e.g., discrete {1,0} or Gaussian representation) is not well correlated with the pose regression quality. 2) The instance feature representation used for predicting the instance score does not explicitly encode the predicted structural pose information, thus is hard to accurately estimate the pose quality score.

To alleviate the above issues, in contrast to previous methods that leverage the discrete {1,0} or Gaussian kernel to denote the confidence of instance localization, we firstly propose to construct the Consistent Instance Representation (CIR) that unifies the pose regression quality score of person instance and the confidence of background into a pixel-wise score map to fill the first gap. Thus, the CIR is able to denote the position of person instance and corresponding pose regression quality simultaneously. Furthermore, to address the second gap, we introduce the Query Encoding Module (QEM) including the Keypoint Query Encoding (KQE) and the Pose Query Encoding (PQE). Specifically, we utilize the Keypoint Query Encoding (KQE) to encode the positional and semantic information of each keypoint for precisely regression, as well as leverage the Pose Query Encoding (PQE) to explicitly encode the predicted structural pose information into instance feature via positional information of each keypoint query. As shown in Figure 1(b), our proposed approach is able to alleviate the inconsistency between instance score and pose regression quality via the proposed Consistent Instance Representation (CIR) and Query Encoding Module (QEM).

The main contributions in our paper can be summarized as follows:

- We propose the Consistent Instance Representation(CIR) that unifies the pose regression quality score of person instance and the presence confidence of background into pixel-wise score map, which alleviates the inconsistency between instance score and pose regression quality.

- We further introduce the Query Encoding Module (QEM) including Keypoint Query Encoding (KQE) to encode the positional and semantic information of each keypoint for precisely regression and Pose Query Encoding(PQE) which explicitly involves the predicted structural pose information into instance feature via the positions of keypoint queries predicted by KQE.

- Based on the proposed CIR and QEM, our approach outperforms previous single-stage even bottom-up methods and achieves the state-of-the-art performance with 71.7 AP on MS COCO test-dev set. To our best knowledge, our paper is the first to attempt to fill the aforementioned gaps in the single-stage pose regression paradigm.

Related Work

In this section, we will review three aspects related to our paper consist of quality estimation, top-down and bottom-up methods as well as single-stage pose regression.

Quality Estimation. Recently, quality estimation is applied in many vision tasks. For instance, IOU-Net (Jiang et al. 2018) adds a parallel branch to predict the IOU between each detected bounding box and the matched ground-truth, which improves the NMS procedure by preserving accurately localized bounding boxes. Mask Scoring RCNN (Huang et al. 2019) proposes to learn the quality of the predicted instance masks via a network block which takes the instance feature and the corresponding predicted mask together to regress the mask IOU. FCOS (Tian et al. 2019) presents a centerness branch to suppress the low-quality detections produced by the locations far away from the object center. The above methods add an additional branch to predict the quality score and the supervision is only assigned for positive samples. Moreover, the classfication score and quality score are multiplied for conducting NMS process. By contrast, we leverage a single branch to learn the proposed Consistent Instance Representation (CIR), which avoids the burden of additional quality estimation branch and the unreliable bias caused by multiplying the instance score with improper quality score.

Top-Down and Bottom-Up Methods. Most existing multi-person pose estimation works can be categorized into top-down and bottom-up methods. Top-down methods (Chen et al. 2018; Fang et al. 2017; Sun et al. 2018, 2019) firstly detect and crop the person region from the image and then predict the single person pose. HRNet (Sun et al. 2019) maintains high-resolution representations and repeatedly aggregates multi-resolution representations to obtain reliable high-resolution representations. SPCNet (Xiao et al. 2020) proposes to incorporate a Dilated Hourglass Module and a Selective Information Module into Hourglass-style network which preserves high spatial resolution and adaptively assembles the multi-level features for detecting the single-person keypoints. Bottom-up methods (Cao et al. 2017; Cheng et al. 2020; Newell, Huang, and Deng 2017) firstly detect the all keypoints in the image and assign the keypoints to individuals via a heuristic grouping process. Associative Embedding (Newell, Huang, and Deng 2017) proposes to predict the keypoint heatmap and tag map simultaneously, and then groups the keypoints with similar tags into individuals. HigherHRNet(Cheng et al. 2020) presents a Higher-Resolution Network to learn high-resolution feature pyramids to better detect the keypoints of the persons with various scales, and follows the AE (Newell, Huang, and Deng 2017) to group the keypoints. Both top-down and bottom-up methods need a additional stage to associate the person instance with corresponding keypoints.

Single-Stage Pose Regression. Recent practices provide a single-stage solution via pixel-wise instance localization and corresponding keypoint regression. They have draw much attention since its compact and end-to-end pipeline. For instance, DirectPose (Tian, Chen, and Shen 2019) proposes a keypoint alignment (KPAlign) module to overcome the misalignment between the features and the predictions. SPM
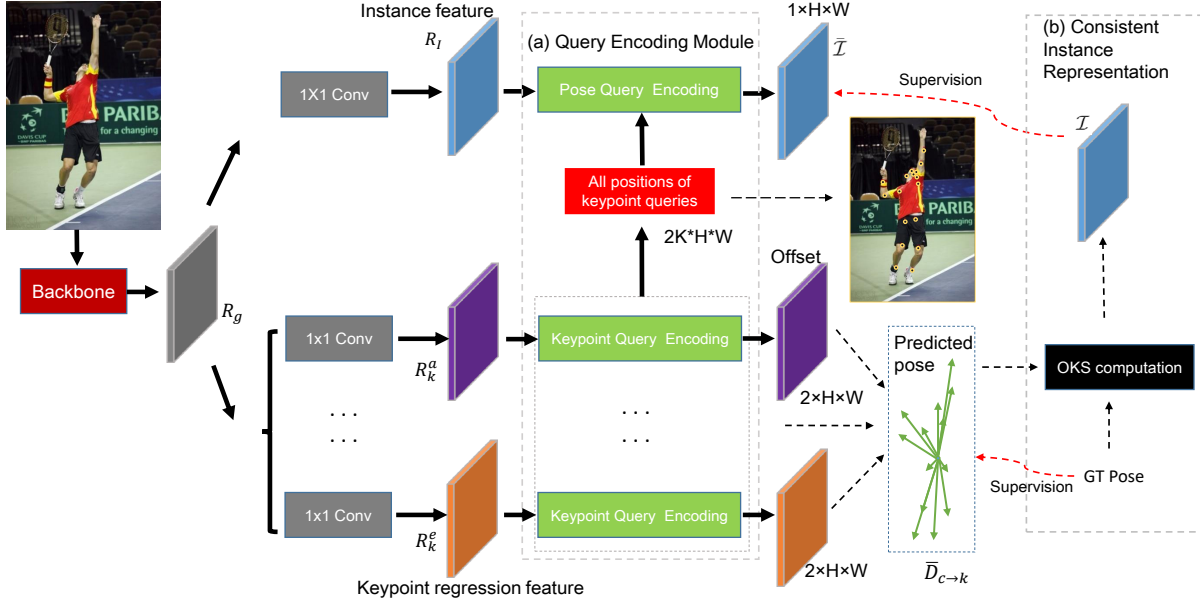
Figure 2: The schematic diagram of our proposed network which uses the multi-branch structure to separately regress the keypoint. Note that R_k represents the different keypoint features (e.g., R_k^a for ankle, R_k^e for elbow). (a) The proposed Query Encoding Module including Keypoint Query Encoding which encodes the position and sematic information of each keypoint for precisely regression and Pose Query Encoding that merges the structured pose information into instance feature. (b) The Consistent Instance Representation that leverages the pose regression quality score to denote the confidence of instance localization. The position of keypoint queries are visualized by yellow circle. K is the number of keypoints queries.

(Nie et al. 2019) proposes a hierarchical Structured Pose Representation according to the body structure to regress accumulative short-range offsets instead of directly regressing long-range offset. PointsetNet (Wei et al. 2020) performs regression from a set of points placed at more advantageous positions which provide informative features and task-specific initializations. However, the above methods concentrates on how to accurately regress the keypoint while ignore the two gaps mentioned above. In this paper, we propose the Consistent Instance Representation and the Query Encoding Module to involve the structural pose information into the instance confidence label and instance feature representation simultaneously to fill the two gaps.

Methods

In this section, we firstly review the formulation of single-stage pixel-wise pose regression. Then, we describe the proposed Consistent Instance Representation (CIR). Finally, we elaborate on the proposed Query Encoding Module (QEM) including Keypoint Query Encoding (KQE) and Pose Query Encoding (PQE).

Single-Stage Pose Regression Formulation

Single-stage multi-person pose estimation methods generally formulate this task as pixel-wise person instance localization and corresponding keypoint regression. It firstly encodes the input image I to produce the general feature representation via the backbone, which is formulated as

$R_g = \mathcal{F}_{backbone}(I) \in \mathbb{R}^{C \times H \times W}$. Then the followed two parallel sub-branch are employed to perform pixel-wise person instance localization and corresponding keypoint offset regression. Concretely, one is to convert the R_g to instance representation R_I, in which each pixel embedding is used to represent an instance and predict the corresponding instance score. The other is to transform R_g to keypoint regression representation R_k, from which the pixel embedding is used to regress the corresponding keypoint displacements. During inference, the regressed pose with high instance score is selected as pose candidates to evaluate the performance.

We consider the above formulation exists two gaps: 1) The instance score is used to select the pose candidates while it is not well correlated with the pose regression quality. 2) The pixel embedding in instance feature R_I does not explicitly encode the corresponding predicted pose information to estimate the reasonable pose regression quality score. Thus we propose the Consistent Instance Representation and Query Encoding Module to attempt to fill the above gaps.

Consistent Instance Representation

To address the inconsistency between the instance score and pose regression quality caused by using the prior instance confidence label (discrete {1,0} and Gaussian kernel) to denote the confidence of instance localization, we construct the Consistent Instance Representation (CIR) that leverages the pose regression quality score to indicate the position and confidence of person instance.
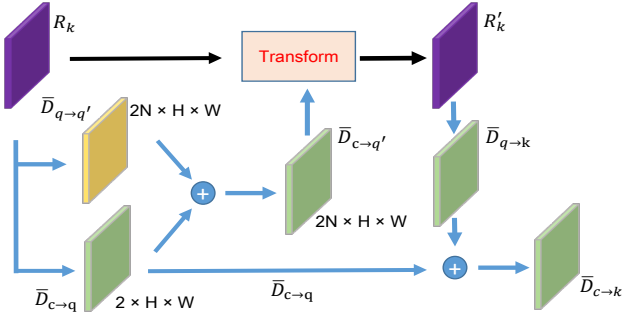
Figure 3: The architecture of Keypoint Query Encoding. Transform refers to extract and aggregate the feature vectors of semantic points q'. N is the number of semantic points.

We leverage Object Keypoint Similarity (OKS) (Lin et al. 2014) between the predicted human pose and the corresponding ground-truth pose to quantify the pose regression quality. In particular, the Object Keypoint Similarity (OKS) of pose $\mathcal{P}$ is formulated as follows:

$$OKS_{\mathcal{P}} = \frac{\sum_i exp(\frac{-d^2_{\mathcal{P},i}}{2s^2_{\mathcal{P}}k^2_i})\delta(\upsilon_{\mathcal{P},i} > 0)}{\sum_i \delta(\upsilon_{\mathcal{P},i} > 0)}, \quad (1)$$

where $d_{\mathcal{P},i}$ is the Euclidean distance between the i-th predicted keypoint location and the ground-truth one of pose $\mathcal{P}$, $\upsilon_{\mathcal{P},i}$ indicates visible or not for the i-th keypoint of pose $\mathcal{P}$, $s_{\mathcal{P}}$ refers to the instance scale of $\mathcal{P}$, and k_i is a constant to control falloff for the i-th keypoints.

The Consistent Instance Representation is a pixel-wise score map denoted as $\mathcal{I}$. $\mathcal{I}(x, y)$ refers to the score at position (x, y), which is formulated as:

$$\mathcal{I}(x,y) = \begin{cases} OKS(\bar{\mathcal{P}}_{(x,y)}, \mathcal{P}_n) & if \ (x,y) \in \Omega_n \\ 0 & else, \end{cases} \quad (2)$$

where $\bar{\mathcal{P}}_{(x,y)}$ is the predicted pose at the position (x,y)$\in$ Ω_n and $\mathcal{P}_n$ refers to ground-truth pose of the n-th human instance, Ω_n is neighboring area around the n-th human instance center (x^c_n, y^c_n), which is formulated as $\{(x,y) \mid \sqrt{[(x,y) - (x^c_n, y^c_n)]^2} < \gamma\}$, γ indicates the radius of neighboring area. The Consistent Instance Representation $\mathcal{I} \in [0,1]$ is able to discriminate person instance and background, in which scalar 0 indicates the confidence of background and the others indicate the person instance position and corresponding pose regression quality score.

As shown in Figure 2(b), we predict a score map $\bar{\mathcal{I}}$ to learn the Consistent Instance Representation $\mathcal{I}$ and employ the pixel-wise weighted L2 loss to penalize the predictions as follows:

$$L_{\mathcal{I}} = W * \| \bar{\mathcal{I}} - \mathcal{I} \|^2, \quad (3)$$

where W denotes the pixel-wise weight map, the weight of the human central area is set to 1 and the background region is 0.1. $*$ refers to pixel-wise multiplication operation.
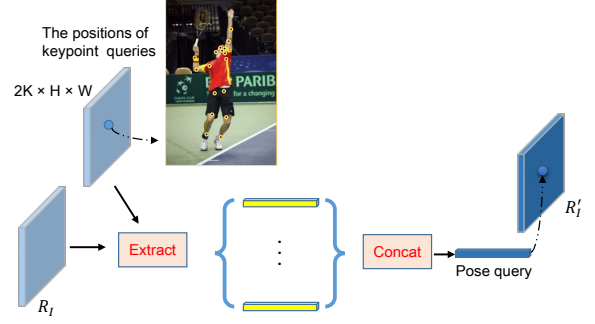

Figure 4: The architecture of Pose Query Encoding.

Query Encoding Module

In order to encode the predicted structural pose information for predicting reasonable pose score, we propose the Query Encoding Module (QEM) including Keypoint Query Encoding (KQE) and Pose Query Encoding (PQE), as shown in Figure2(a). Keypoint Query Encoding (KQE) is presented to encode the positional and semantic information of each keypoint. Pose Query Encoding (PQE) is introduced to encode the structural pose information into instance feature via the each keypoint query's location predicted by KQE.

Keypoint Query Encoding. We use multi-branch structure to perform separated keypoint regression, in which each branch follows the same design. We leverage the general feature R_g to produce the separated keypoint representation R_k (e.g. nose representation R^n_k, ankle representation R^a_k, elbow representation R^e_k, wrist representation R^w_k) via the separated 1×1 convolutional layer and define a keypoint query for each keypoint, which encodes the positional and semantic information to precisely localize the keypoints.

Specifically, we take a pixel position c of instance central area as example to illustrate the Keypoint Query Encoding. The whole process is decomposed into three steps. As shown in Figure 3, first, we predict a displacement $\bar{\mathbf{D}}_{c \to q}$ from center c to dynamically locate the keypoint query q via the separated keypoint representation R_k. Second, owing to keypoint query is a single position and thus limited to encode the sufficient semantic information of corresponding keypoint, we further extract and aggregate the features of N points (named semantic point q') to enrich the semantic information of keypoint query via bilinear interpolation. The semantic points q' are located by regressing N displacements $\bar{\mathbf{D}}_{q \to q'}$ based on the position of keypoint query q. Obtaining the transformed feature R'_k is formualted as : $R'_k(c) = \sum_{n=1}^{N}\{R_k(c + \bar{\mathbf{D}}_{c \to q} + \bar{\mathbf{D}}_{q \to q'_n})\}$, where N is the number of semantic points. Thus, the transformed feature R'_k is considered to encode the sufficient positional and semantic information for each keypoint. Finally, we regress the displacements $\bar{\mathbf{D}}_{q \to k}$ from the keypoint query to corresponding keypoint via the transformed feature R'_k to precisely locate the keypoint. The displacements $\bar{\mathbf{D}}_{c \to k}$ from center to keypoints is formulated as:

$$\bar{\mathbf{D}}^i_{c \to k} = \bar{\mathbf{D}}^i_{c \to q} + \bar{\mathbf{D}}^i_{q \to k}, \quad (4)$$

Methods	Params	Input size	GFLOPs	AP	AP_{50}	AP_M	AP_L
Personlab (Papandreou et al. 2018)	68.7	1401	405.5	66.5	86.2	62.3	73.2
PifPaf (Kreiss, Bertoni, and Alahi 2019)	-	-	-	67.4	-	-	-
CenterNet-HG (Zhou, Wang, and Krähenbühl 2019)	-	512	-	64.0	-	-	-
HrHRNet-W32(Cheng et al. 2020)	28.5	512	47.9	67.1	86.2	-	-
HrHRNet-W48(Cheng et al. 2020)	63.8	640	154.3	69.9	87.2	-	-
DEKR-W32(Geng et al. 2021)	29.6	512	45.4	68.0	86.7	62.1	77.7
DEKR-W48(Geng et al. 2021)	65.7	640	141.5	71.0	88.3	66.7	78.5
Ours (HRNet-W32)	29.7	512	46.4	69.8	88.1	63.8	78.9
Ours (HRNet-W48)	65.8	640	143.4	**72.4**	**89.1**	**67.3**	**80.4**

Table 1: Comparisons with the previous state-of-the-art methods on the COCO mini-val set (single-scale testing).

where i refers to the i-th keypoint of human instance.

We construct a pixel-wise dense offset map $\mathbf{D}_{c \to k}$ as ground-truth to penalize the predicted $\bar{\mathbf{D}}_{c \to k}$. $\mathbf{D}_{c \to k}^i$ indicates the x-y offsets for the i-th keypoints of the person, takes the following form:

$$\mathbf{D}_{c \to k}^i(x,y) = \begin{cases} (x,y) - (x_n^i, y_n^i) & if \ (x,y) \in \Omega_n \\ 0 & else, \end{cases}$$

(5)

where Ω_n is adjacent area around the n-th instance center (x_n^c, y_n^c), which is formulated as $\{(x,y) \mid \sqrt{[(x,y) - (x_n^c, y_n^c)]^2} < \gamma\}$. (x_n^i, y_n^i) indicates the coordinates of the i-th keypoint for the n-th person. The loss function is smooth L1 as follows:

$$L_{\mathbf{D}_{c \to k}} = SmoothL1(\mathbf{D}_{c \to k}, \bar{\mathbf{D}}_{c \to k}). \quad (6)$$

Pose Query Encoding. Based on the keypoint queries, we introduce Pose Query Encoding to involve the predicted structural pose information into instance feature, which is used to predict a pose regression quality score. Concretely, we convert the general feature R_g to produce the raw instance representation R_I, then concatenate the feature vector at the position of all keypoint queries, as illustrated in Figure 4. The above process transforms the raw instance representation R_I to generate the new instance representation R_I', which is formulated as follows:

$$R_I'(c) = Concat(\{R_I(c + \bar{\mathbf{D}}_{c \to q}^i)\}_{i=1}^K), \quad (7)$$

where K is the number of keypoint queries, $\bar{\mathbf{D}}_{c \to q}^i$ is the displacement from center c to i-th keypoint query. Due to each keypoint query encode the positional and semantic information for corresponding keypoint. Thus, the Pose Query Encoding is capable of encoding the predicted structural pose information into instance feature representation. Finally, we leverage the new instance representation R_I' to predict the pixel-wise score map $\bar{\mathcal{I}}$ to better fit the Consistent Instance Representation (CIR).

Experiments

In this section, we first briefly introduce our experimental setup. Then we carry out the ablation study to investigate the effectiveness of each components of our proposed network. Finally, we conduct the comprehensive comparisons

with previous state-of-the-art methods to verify the superiority of our proposed network.

Experimental Setup

Dataset. We conduct our experiments on widely-used pose estimation benchmark MS COCO (Lin et al. 2014), which includes 200k images with 250k human instance annotated with the positions of 17 body joints. Following previous settings, we leverage coco train2017 with 57k images for training, mini-val set with 5k images for conducting ablation studies, test-dev set with 20k images for comparing with the previous state-of-the-art methods.

Evaluation Metric. The evaluation metrics are average precision and average recall scores based on different Object Keypoint Similarity (OKS) thresholds from 0.5 to 0.95 to evaluate the performance.

Augmentation. In training stage, we carry out data augmentation via random flip with probability of 0.5, random rotation in $[-30, 30]$ degrees, random scaling of $[0.75, 1.5]$, random shift of $[-40, 40]$ pixels as well as color jitter to augment training samples. Each input is cropped to 512 / 640 pixels. The output size is 1/4 of the input resolution.

Implementation Details. We train our proposed network via Adam optimizer with a mini-batch size of 64. The initial learning rate is set as 5e-4 and dropped to 5e-5 and 5e-6 at the 150th and 170th epochs respectively. Furthermore, the radius of center-neighboring area γ is set to 4. The loss weight of $L_{\mathcal{I}}$ and $L_{\mathbf{D}_{c \to k}}$ are both set to 1.0. For inference, we keep the aspect ratio of raw input image and resize the short side of the images to 512 / 640 pixels.

Ablative Analysis

In this subsection, we first report the contributions of each component in our framework. Then, we delve into the design of them. All ablation studies adopt HRNet-W32 as backbone with single-scale testing on the COCO mini-val set.

Contributions of each components. We analyze the contribution of each component in our proposed method. The results are shown in Table 2. Note that we adopt the Gaussian kernel as instance confidence label to denote the person instance in Expt.1, 4. With only Consistent Instance Representation (CIR) applied, we achieve 1.0 AP improvements as reported in Expt.1 and Expt.2. Keypoint Query Encoding (KQE) improves 3.9 AP based on CIR as shown in Expt.2 and Expt.3. Pose Query Encoding (PQE) is able to obtain

Expt.	CIR	KQE	PQE	AP	AP_{50}	AP_M	AP_L
1	-	-	-	63.5	85.5	56.5	73.8
2	√	-	-	64.5	85.8	58.0	74.6
3	√	√	-	68.4	87.1	62.2	77.5
4	-	√	√	68.3	86.9	62.0	77.3
5	√	√	√	**69.8**	**88.1**	**63.8**	**78.9**

Table 2: Contributions of each components. CIR denotes the Consistent Instance Representation. KQE is the Keypoint Query Encoding. PQE indicates the Pose Query Encoding.

Instance Rep.	QEB	AP	AP_{50}	AP_M	AP_L
Discrete $\{0,1\}$	-	67.7	86.7	61.4	77.2
Discrete $\{0,1\}$	√	68.6	87.2	62.5	77.9
Gaussian	-	68.3	86.9	62.0	77.3
Gaussian	√	69.0	87.6	63.1	78.3
CIR(ours)	-	**69.8**	**88.1**	**63.8**	**78.9**

Table 3: Comparisons with previous instance confidence representation for denoting the confidence of instance localization. QEB refers to employ an additional quality estimation branch to modulate the instance score.

1.4 AP improvements in Expt.3 and Expt.5. Furthermore, as shown in Expt.4 and Expt.5, based on the KQE and PQE, the Consistent Instance Representation (CIR) is capable of improving 1.5 AP. The exhaustive analysis for each components will be described in below.

Analysis of Consistent Instance Representation. The previous studies (Tian, Chen, and Shen 2019; Nie et al. 2019) employ discrete $\{1,0\}$ or the 2-dimensional Gaussian kernel to denote the confidence of instance localization. The former assigns the pixels of human central area with label $\{1\}$. This scheme treats each pixel of central area equally while ignores the difference of pose regression quality. The later hypothetically considers that the center position will predict the optimal pose. The farther away from the center, the worse the pose regression quality is. However, both of them are manually settled and may result in the gap between instance score and corresponding pose regression quality.

For the previous two instance confidence labels, we further add an additional quality estimation branch, following (Jiang et al. 2018; Huang et al. 2019), whose supervision is only assigned for instance area. The estimated instance score and pose regression quality score are multiplied to select pose candidates during inference. We construct 4 contrasts including discrete $\{1,0\}$ without or with additional quality estimation branch, Gaussian distribution without or with additional quality estimation branch. Our CIR unifies the pose regression quality score of instance area and confidence of background into a single pixel-wise score map thus the additional quality estimation branch is no longer required. As shown in Table 3, we observe that the additional quality estimation branch is able to consistently improve the performance for both discrete $\{1,0\}$ and Gaussian representation. However, as shown in Figure 5, we observe that employing an additional quality estimation branch to modulate the instance score may lead to false positives (the blue points in

N	0	3	6	9	12	15
AP	68.5	68.9	69.2	**69.8**	69.6	69.1

Table 4: Ablation study for varying the number of semantic points N by fixing the other proposed components.
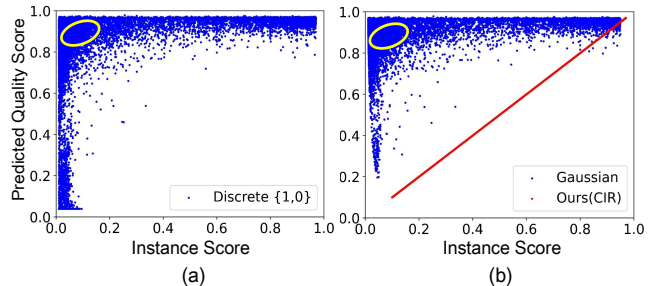

Figure 5: The blue points in scatter diagram (a) and (b): employing an additional quality estimation branch to modulate instance score supervised by discrete $\{1,0\}$ or Gaussian representation. The points denote the randomly remained pose candidates after NMS (5 per image) with the predicted instance score and pose quality score. The blue points in yellow circles indicate the pose candidates with low instance score and high pose quality score. Some of these may predicted by background area which lead to false positives, due to the supervision of additional quality estimation is only assigned for instance area during training. The red points in scatter diagram (b): our CIR avoids the unreliable bias caused by multiplying the instance score and uncontrollably high quality score, thus achieves the better performance.

yellow circles). Due to the supervision of quality estimation is only assigned for instance area, thus the network may predicts the uncontrollably high score in background to raise the corresponding low instance score. Our proposed Consistent Instance Representation unifies the pose regression quality score of instance area and confidence of background into a single pixel-wise map which avoids the unreliable bias caused by multiplying the instance score and uncontrollably quality score in background. As a result, our CIR achieves the better performance.

Analysis of Keypoint Query Encoding. The semantic information for each keypoint query is supplemented via the features of N semantic points. We explore the influence for the number of semantic points in Keypoint Query Encoding. As shown in Table 4, our method achieves the better performance with the number of semantic point increasing. It proves that enriching the semantic feature for keypoint query is able to more precisely regress the keypoint. We achieve the best performance with 69.8 AP when N is set to 9.

Analysis of Pose Query Encoding. Pose Query Encoding aims to encode the predicted structural pose information into instance feature to predict pose regression quality score. Based on the other proposed components, we investigate how to generate the pose query that better involve the predicted pose information into instance feature. As reported in

Methods	Input size	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR
Bottom-Up Methods							
CMU-Pose*† (Cao et al. 2017)	-	61.8	84.9	67.5	57.1	68.2	66.5
AE*† (Newell, Huang, and Deng 2017)	512	65.5	86.8	72.3	60.6	72.6	70.2
CenterNet-DLA (Zhou, Wang, and Krähenbühl 2019)	512	57.9	84.7	63.1	52.5	67.4	-
CenterNet-HG (Zhou, Wang, and Krähenbühl 2019)	512	63.0	86.8	69.6	58.9	70.4	-
PifPaf (Kreiss, Bertoni, and Alahi 2019)	801	66.7	-	-	62.4	72.9	-
HrHRNet-w48* (Cheng et al. 2020)	512	68.4	88.2	75.1	64.4	74.2	-
SWAHR(HrHRNet-W48)* (Luo et al. 2021)	640	70.2	89.9	76.9	65.2	77.0	-
Single-Stage Regression Methods							
SPM *† (Nie et al. 2019)	-	66.9	88.5	72.9	62.6	73.1	-
DirectPose † (Tian, Chen, and Shen 2019)	800*1333	64.8	87.8	71.1	60.4	71.5	-
MDN₃ (Varamesh and Tuytelaars 2020)	-	62.9	85.1	69.4	58.8	71.4	-
PointSetNet *† (Wei et al. 2020)	640	68.7	89.9	76.3	64.8	75.3	74.8
DEkR-W48*(Geng et al. 2021)	640	70.0	89.4	77.3	65.7	76.9	75.4
DEkR-W48 *†(Geng et al. 2021)	640	71.0	89.2	78.0	67.1	76.9	**76.7**
Ours (HRNet-W32)	512	69.0	89.3	76.0	62.8	77.0	73.6
Ours (HRNet-W32)†	512	70.5	89.6	77.5	64.8	78.0	75.1
Ours (HRNet-W48)	640	71.0	90.2	78.2	66.2	77.8	76.0
Ours (HRNet-W48)†	640	**71.7**	**90.4**	**78.7**	**67.3**	**78.5**	76.5

Table 5: Comparisons with the state-of-the-art methods on COCO test-dev set. ∗ indicates using additional post-process(e.g., single-person model refinement used in CMU-Pose, AE, SPM and pose scoring net in DEKR). † refers to multi-scale testing.

Method	AP	AP_{50}	AP_{75}	AP_M	AP_L
Auto	68.8	87.6	75.1	62.3	77.9
SP	68.9	87.8	75.0	62.5	78.3
KPS	69.0	87.5	75.1	62.6	78.5
KQ (ours)	**69.8**	**88.1**	**76.2**	**63.8**	**78.9**

Table 6: Ablation study for the construction of pose query. Auto: aggregating the features via the positions of 17 automatically located points; SP: aggregating the features via the positions of 17*N semantic points; KPS: aggregating the features via the position of 17 keypoints; KQ: aggregating the features via the position of 17 keypoint queries.

Table 6, we construct 4 contrasts for comparisons, the results prove that aggregating the features via the positions of 17 keypoint queries to construct the pose query achieves the better performance compared with the others.

Comparisons

Mini-val Results. Table 1 reports the performance of single-scale testing on COCO mini-val set. With HRNet-W32 as backbone, our method achieves 69.8 AP when input resolution is set as 512 pixels and outperforms the previous bottom-up methods (Papandreou et al. 2018; Zhou, Wang, and Krähenbühl 2019) with a large margin. In particular, compared with the state-of-the-art HigherHRNet (Cheng et al. 2020) and DEKR(Geng et al. 2021), our network achieves 2.7 AP and 1.8 AP improvements without either multi-scale heatmap aggregation or additional pose scoring net during inference. We further obtain 72.4 AP with input resolution of 640 pixels via HRNet-W48, which is a new state-of-the-art performance compared with all existing single-stage as well as bottom-up methods.
Test-dev Results. Table 5 reports the comparisons with

the prior bottom-up and single-stage regression-based methods on test-dev2017 set. Adopting HRNet-W48 as backbone with single-scale testing, our method achieves 71.0 AP which outperforms the bottom-up AE (Newell, Huang, and Deng 2017), CenterNet-HG (Zhou, Wang, and Krähenbühl 2019) as well as PifPaf (Kreiss, Bertoni, and Alahi 2019) with a large margin, and surpasses the state-of-the-art HigherHRNet-W48 + AE (Cheng et al. 2020) and SWAHR-W48 (Luo et al. 2021) by 2.6 AP and 0.8 AP respectively. Compared with single-stage regression-based methods, our approach achieves 4.8 AP gains over SPM (Nie et al. 2019), 6.9 AP gains over Directpose (Tian, Chen, and Shen 2019) and 3.0 AP improvements over PointsetNet (Wei et al. 2020). Moreover, we improve 1.0 AP and 0.7 AP in comparison to state-of-the-art regression-based model DEKR-W48 (Geng et al. 2021) for single-scale and multi-scale testing.

Conclusion

In this paper, we propose to learn the regression quality-aware representation. Concretely, we firstly present the Consistent Instance Representation (CIR) that unifies the pose regression quality score of instance area and presence confidence of background into a single pixel-wise score map to fill the inconsistency between the instance score and pose regression quality. Furthermore, we present the Query Encoding Module (QEM) that encodes the keypoint information for precisely regression and involves the predicted pose information into instance feature representation to predict the reasonable pose regression quality score. With the proposed CIR and QEM, our network is able to significantly alleviate the above gaps existing in current single-stage pose regression practices. Comprehensive experiments demonstrate the state-of-the-art performance of our proposed method.

Acknowledgements

References

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7103–7112.

Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; and Zhang, L. 2020. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5386–5395.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2334–2343.

Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; and Wang, J. 2021. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14676–14686.

Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; and Wang, X. 2019. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6409–6418.

Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, 784–799.

Kreiss, S.; Bertoni, L.; and Alahi, A. 2019. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11977–11986.

Law, H.; and Deng, J. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 734–750.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3595–3603.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Luo, Z.; Wang, Z.; Huang, Y.; Wang, L.; Tan, T.; and Zhou, E. 2021. Rethinking the Heatmap Regression for Bottom-up Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13264–13273.

Newell, A.; Huang, Z.; and Deng, J. 2017. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in neural information processing systems*, 2277–2287.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, 483–499. Springer.

Nie, X.; Feng, J.; Zhang, J.; and Yan, S. 2019. Single-stage multi-person pose machines. In *Proceedings of the IEEE International Conference on Computer Vision*, 6951–6960.

Papandreou, G.; Zhu, T.; Chen, L.-C.; Gidaris, S.; Tompson, J.; and Murphy, K. 2018. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 269–286.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12026–12035.

Su, K.; Yu, D.; Xu, Z.; Geng, X.; and Wang, C. 2019. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5674–5682.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5693–5703.

Sun, X.; Xiao, B.; Wei, F.; Liang, S.; and Wei, Y. 2018. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 529–545.

Tian, Z.; Chen, H.; and Shen, C. 2019. DirectPose: Direct End-to-End Multi-Person Pose Estimation. *arXiv preprint arXiv:1911.07451*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, 9627–9636.

Varamesh, A.; and Tuytelaars, T. 2020. Mixture dense regression for object detection and human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13086–13095.

Wei, F.; Sun, X.; Li, H.; Wang, J.; and Lin, S. 2020. Point-set anchors for object detection, instance segmentation and pose

estimation. In *European Conference on Computer Vision*, 527–544. Springer.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, 466–481.

Xiao, Y.; Yu, D.; Wang, X.; Lv, T.; Fan, Y.; and Wu, L. 2020. SPCNet:Spatial Preserve and Content-aware Network for Human Pose Estimation. In *European Conference on Artificial Intelligence*.

Yu, D.; Su, K.; Sun, J.; and Wang, C. 2018a. Multi-person Pose Estimation for Pose Tracking with Enhanced Cascaded Pyramid Network. In *European Conference on Computer Vision*.

Yu, F.; Wang, D.; Shelhamer, E.; and Darrell, T. 2018b. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2403–2412.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as points. *arXiv preprint arXiv:1904.07850*.